Reference Document AK

Requested Patent: JP2000172614A

Title: INTERNET RETRIEVING APPARATUS ;

Abstracted Patent: JP2000172614 ;

Publication Date: 2000-06-23 ;

Inventor(s): YANAGIDA KEITON ;

Applicant(s): NEC CORP ;

Application Number: JP19980351931 19981210 ;

Priority Number(s): ;

IPC Classification: G06F13/00 ; G06F17/30 ;

Equivalents:

ABSTRACT:

PROBLEM TO BE SOLVED: To automatically retrieve and find out the uniform resource locator(URL) of a changed web page. SOLUTION: This apparatus 10 is composed of a web page retrieving part 11, link destination retrieving part 12, link destination contents reading part 13, read contents storage part 14, disk cache 15, disk cache internal contents comparing part 16 and compared result reporting part 17. A URL corresponding to the master of the web page desired to connect is retrieved, the URL of the link destination is searched from the retrieved web page, and the contents of the web page at the link destination are read and stored in the read contents storage part 14. Moreover, the contents before the web page desired to connect are compared with the contents of the link destination stored in the read content storage part 14 and with the rate of coincidence as a reference, it is assessed whether or not the contents of both the web pages are equal.

BEST AVAILABLE COPY

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－172614
（Ｐ2000－172614Ａ）
(43)公開日　平成12年６月23日(2000.6.23)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'(参考)
Ｇ06Ｆ 13/00	354	Ｇ06Ｆ 13/00	354Ｄ	5Ｂ075
17/30		15/40	310Ｃ	5Ｂ089

審査請求 有　請求項の数2 ＯＬ （全 4 頁）

(21)出願番号　特願平10－351931

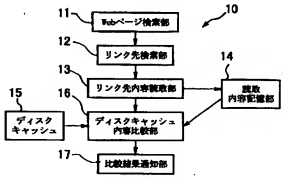
(22)出願日　平成10年12月10日(1998.12.10)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　柳田 直敦
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100108578
弁理士　高橋 詔男　（外３名）

Ｆターム(参考)　5Ｂ075 ＫＫ02 ＮＤ02 ＮＤ06 ＮＤ35 ＮＫ10
ＰＲ06 ＱＭ08
5Ｂ089 ＧＡ12 ＧＡ21 ＧＢ04 ＪＡ24 ＫＡ03
ＬＢ04 ＬＢ14 ＬＢ25

(54)【発明の名称】　インターネット検索装置

(57)【要約】
【課題】　変更されたＷｅｂページのＵＲＬを自動的に検索して見つける。
【解決手段】　インターネット検索装置１０は、Ｗｅｂページ検索部１１と、リンク先検索部１２と、リンク先内容読取部１３と、読取内容記憶部１４と、ディスクキャッシュ１５と、ディスクキャッシュ内容比較部１６と、比較結果通知部１７とから構成した。接続を希望するＷｅｂページの親に相当するＵＲＬを検索し、検索されたＷｅｂページからリンク先のＵＲＬを探し出し、リンク先のＷｅｂページの内容を読み取り、読取内容記憶部１４に格納する。また、接続を希望するＷｅｂページの以前の内容と、読取内容記憶部１４に格納されたリンク先の内容とを比較し、両Ｗｅｂページの内容が同一であるか否かを一致率を基準として判断する。



【特許請求の範囲】

【請求項1】 World Wide Webにおいて、階層構造をなす複数のＷｅｂページから接続を希望するＷｅｂページを検索するインターネット検索装置であって、

前記インターネット検索装置は、接続に失敗したＷｅｂページのＵＲＬよりも上位層のＷｅｂページのＵＲＬを推定して検索する手段と、前記上位層のＷｅｂページからリンク先を抽出する手段と、この抽出されたリンク先のＷｅｂページの内容を読み取って記憶する手段と、予め前記接続に失敗したＷｅｂページの以前の接続時における内容を記憶する手段と、この以前の接続時における内容と前記リンク先のＷｅｂページの内容とを比較して同一であるか否かを判断する手段とを備えていることを特徴とするインターネット検索装置。

【請求項2】 前記インターネット検索装置は、前記接続に失敗したＷｅｂページの以前の接続時における内容と、前記リンク先のＷｅｂページの内容とを比較する際に、両者の一致率が所定の一致率を越えたときに両者が同一のＷｅｂページであると判断し、前記両者の一致率が前記所定の一致率に満たないときに両者は異なるＷｅｂページであると判断する手段を備えていることを特徴とする請求項1に記載のインターネット検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、World Wide Webにおけるウェブページを検索するインターネット検索装置に関する。

【０００２】

【従来の技術】WWW（World Wide Web）はネットワーク上に分散された多数のＨＴＴＰ（Hyper Text Transfer Protocol）サーバにより提供されるＷｅｂページが相互にリンクされたもので、ＷｅｂページはＨＴＭＬ（Hyper Text Markup Languages）で記述されたテキストファイルと画像データ等から構成されている。クライアントとなるユーザはＷｅｂブラウザ等を使用することでＨＴＴＰサーバに接続し、Ｗｅｂページのデータをダウンロードしてウェブページを表示する。各Ｗｅｂページはネットワーク上における所在がＵＲＬ（Uniform Resource Locator、最近はＵＲＩ：Uniformed Resource Identifierと呼ばれる。）によって指定されており、ユーザはＷｅｂブラウザ等において接続を希望するＷｅｂページのＵＲＬを入力することによって所望のＷｅｂページに接続することができる。ところで従来においては、例えばＷｅｂページの内容が何ら変更されていなくても、このＷｅｂページのＵＲＬが変更されて、移動先のＵＲＬが提示されていない場合においては、ユーザはこのＷｅｂページに接続することができなくなり、移動先のＵＲＬを検索システム等を使用して探し出す必要が生じていた。

【０００３】

【発明が解決しようとする課題】上記のように、ＵＲＬが変更されたＷｅｂページを検索システム等を利用して検索する場合、一般に、検索用に入力するキーワードに対して複数のＷｅｂページの候補が存在するが、特に多量のＷｅｂページが候補として提示された場合には、目的とするＷｅｂページを見つけ出すまでに多くの時間と手間が必要になるという問題が生じる。また、接続を希望するＷｅｂページのＵＲＬが変更された直後においては、検索システム等に登録されている内容が更新されていないおそれがあり、移動先のＵＲＬを見つけ出すことが困難になるという問題が生じる。本発明は上記事情に鑑みてなされたもので、接続を希望するＷｅｂページのＵＲＬが変更された際にも、変更後のＵＲＬを自動的に検出することが可能なインターネット検索装置を提供することを目的とする。

【０００４】

【課題を解決するための手段】上記課題を解決して係る目的を達成するために、請求項1に記載の本発明のインターネット検索装置は、World Wide Webにおいて、階層構造をなす複数のＷｅｂページから接続を希望するＷｅｂページを検索するインターネット検索装置であって、前記インターネット検索装置は、接続に失敗したＷｅｂページのＵＲＬよりも上位層のＷｅｂページのＵＲＬを推定して検索する手段と、前記上位層のＷｅｂページからリンク先を抽出する手段と、この抽出されたリンク先のＷｅｂページの内容を読み取って記憶する手段と、予め前記接続に失敗したＷｅｂページの以前の接続時における内容を記憶する手段と、この以前の接続時における内容と前記リンク先のＷｅｂページの内容とを比較して同一であるか否かを判断する手段とを備えていることを特徴としている。

【０００５】上記構成のインターネット検索装置では、例えばＨＴＴＰサーバ側においてＷｅｂページの構成等を変更したことに伴って接続を希望するＷｅｂページのＵＲＬが変更され、クライアントとされるユーザから接続を希望するＷｅｂページへの接続ができなくなった際にも、接続を希望するＷｅｂページに比べてＵＲＬが変更される可能性の低い上位層のＷｅｂページに接続することによって、接続を希望するＷｅｂページに設定された新たなＵＲＬを探し出すことが可能である。

【０００６】さらに、請求項2に記載のインターネット検索装置は、前記インターネット検索装置は、前記接続に失敗したＷｅｂページの以前の接続時における内容と、前記リンク先のＷｅｂページの内容とを比較する際に、両者の一致率が所定の一致率を越えたときに両者が同一のＷｅｂページであると判断し、前記両者の一致率が前記所定の一致率に満たないときに両者は異なるＷｅｂページであると判断する手段を備えていることを特徴としている。

【０００７】上記のようなインターネット検索装置で

は、接続を希望するＷｅｂページのＵＲＬが変更されたことに加えて、接続を希望するＷｅｂページの内容が更新された場合であっても、この更新されたＷｅｂページの内容と、以前に接続した時点におけるＷｅｂページの内容との比較をおこなう際に、一致率を適宜に設定することによって、例えば両者の内容が完全に一致していなくても、両者が同一のＷｅｂページであると判断することができるため、接続を希望するＷｅｂページの変更後のＵＲＬを見つけることが可能である。

【０００８】

【発明の実施の形態】以下、本発明のインターネット検索装置の実施形態について添付図面を参照しながら説明する。図１は本発明の一実施形態に係わるインターネット検索装置を示すブロック構成図である。本実施の形態によるインターネット検索装置１０は、Ｗｅｂページ検索部１１と、リンク先検索部１２と、リンク先内容読取部１３と、読取内容記憶部１４と、ディスクキャッシュ１５と、ディスクキャッシュ内容比較部１６と、比較結果通知部１７とから構成されている。Ｗｅｂページ検索部１１は、例えば接続を希望するＷｅｂページの親（上位層）に相当するＷｅｂページのＵＲＬ、すなわち例えば接続希望のＵＲＬがhttp://www.nec.co.jp/usr/index.htmlである場合に、このＵＲＬの親に相当するＵＲＬとして例えばhttp://www.nec.co.jp/index.html等のＷｅｂページを検索する。 リンク先検索部１２は、Ｗｅｂページ検索部１１で検索されたＷｅｂページに対して、このＷｅｂページからリンクされているリンク先のＷｅｂページのＵＲＬを探し出す。リンク先内容読取部１３は、リンク先のＷｅｂページの内容を読み取り、読取内容記憶部１４に格納する。ディスクキャッシュ１５は、予め、以前に接続した時点での接続希望のＷｅｂページの内容を記憶しておく。ディスクキャッシュ内容比較部１６は、ディスクキャッシュ１５に記憶されている接続を希望するＷｅｂページの以前の内容と、リンク先内容読取部１３で読み込まれて読取内容記憶部１４に格納されたリンク先のＷｅｂページの内容とを比較し、両Ｗｅｂページの内容が同一であるか否かを、予め設定される例えば一致率等を基準として判断する。比較結果通知部１７は、ディスクキャッシュ内容比較部１６での比較結果を通知する。

【０００９】本実施の形態によるインターネット検索装置１０は上述の構成を備えており、次に、インターネット検索装置１０の動作について図１および図２を参照しながら説明する。図２は図１に示すインターネット検索装置１０の動作を示すフローチャートである。先ず、ユーザが例えばＷｅｂブラウザ等を使用して接続を希望するＷｅｂページのＵＲＬに接続を試みた時に、例えば該当するＷｅｂページが見つからない等のエラーがＨＴＴＰサーバ側から通知されると、インターネット検索装置１０が起動する（ステップＳ１）。インターネット検索装置１０のＷｅｂページ検索部１１は、接続に失敗したＷｅｂページのＵＲＬに基づいて、このＷｅｂページの親に相当するＷｅｂページのＵＲＬを推定して接続を行う（ステップＳ２）。ここで、例えば接続希望のＵＲＬがhttp://www.nec.co.jp/usr/index.htmlである場合には、このＵＲＬの親に相当するＵＲＬとして例えばhttp://www.nec.co.jp/index.htmlを採用する。

【００１０】親に相当するＷｅｂページに接続すると、リンク先検索部１２は、このＷｅｂページからリンクされているリンク先のＷｅｂページのＵＲＬを探して接続を行い（ステップＳ３）、リンク先内容読取部１３がリンク先のＷｅｂページの内容を読み取って読取内容記憶部１４に格納する（ステップＳ４）。一方、ディスクキャッシュ１５には、接続を希望するＷｅｂページの以前の内容、すなわちＵＲＬが変更される前に接続した際に記憶されたＷｅｂページの内容が格納されている。そこで、ディスクキャッシュ内容比較部１６は、ディスクキャッシュ１５に記憶されている接続を希望するＷｅｂページの以前の内容と、リンク先内容読取部１３で読み込まれて読取内容記憶部１４に格納されたリンク先のＷｅｂページの内容とを比較し、所定の例えば５０％以上の一致率が得られた際に両Ｗｅｂページの内容が同一であると判断し、一方、例えば５０％未満の一致率では互いに異なるＷｅｂページであると判断する（ステップＳ５）。比較結果通知部１７は、ディスクキャッシュ内容比較部１６において両Ｗｅｂページが同一であると判断された際には、リンク先内容読取部１３が読み取ったリンク先のＷｅｂページのＵＲＬをユーザに通知し、逆に両Ｗｅｂページが互いに異なると判断された際には、ＵＲＬの検出に失敗した旨の通知をユーザに行う（ステップＳ６）。

【００１１】本実施の形態によるインターネット検索装置１０によれば、例えばＨＴＴＰサーバ側においてＷｅｂページの構成等を変更したことに伴って接続を希望するＷｅｂページのＵＲＬが変更され、クライアントとされるユーザから接続を希望するＷｅｂページへの接続ができなくなった際にも、接続を希望するＷｅｂページに比べてＵＲＬが変更される可能性が低い親のＷｅｂページに接続することによって、接続を希望するＷｅｂページに設定された新たなＵＲＬを探し出すことが可能である。また、例えばＷｅｂページの構成等に加えて、接続を希望するＷｅｂページの内容が更新された場合であっても、以前に接続した時点でのＷｅｂページの内容と、更新されたＷｅｂページの内容との比較の際に、適宜の一致率を基準値として設定することによって、両Ｗｅｂページの内容が完全に一致していなくても、両者が同一のＷｅｂページであると判断することができるため、接続を希望するＷｅｂページの変更後のＵＲＬを見つけることが可能である。

【００１２】なお、本実施の形態においては、接続に失

敗したＷｅｂページの親に相当するＷｅｂページのＵＲＬを推定するとしたが、これに限定されず、例えば同じＨＴＴＰサーバ上のＵＲＬであって、ＵＲＬの末尾が例えばｉｎｄｅｘ．ｈｔｍｌとなっているＷｅｂページを検索してもよい。ここで、例えばｉｎｄｅｘ．ｈｔｍｌは、一般に、ＨＴＴＰサーバの標準設定によって作成されるＷｅｂページであって、Ｗｅｂページの構成が変更される際にもＵＲＬが消去される可能性が低い。要するに、接続の失敗したＷｅｂページのＵＲＬの近隣のＵＲＬであって、確実に接続が可能なＵＲＬを推定すればよい。

【００１３】

【発明の効果】以上説明したように、請求項１記載の本発明のインターネット検索装置によれば、接続に失敗したＷｅｂページに比べてＵＲＬが変更される可能性の低い上位層のＷｅｂページに接続することによって、接続に失敗したＷｅｂページに設定された新たなＵＲＬを探し出すことが可能である。さらに、請求項２記載のインターネット検索装置によれば、接続を希望するＷｅｂページの内容が更新された場合であっても、以前に接続した時点でのＷｅｂページの内容と、更新されたＷｅｂページの内容との比較の際に、適宜の一致率を基準値として設定することによって、両Ｗｅｂページの内容が完全に一致していなくても、接続を希望するＷｅｂページの変更後のＵＲＬを見つけることが可能である。

【図面の簡単な説明】

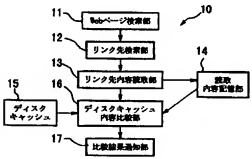
【図１】　本発明の一実施形態に係わるインターネット検索装置のブロック構成図である。

【図２】　図１に示すインターネット検索装置の動作を示すフローチャートである。

【符号の説明】

１０　インターネット検索装置
１１　Ｗｅｂページ検索部
１２　リンク先検索部
１３　リンク先内容読取部
１４　読取内容記憶部
１５　ディスクキャッシュ
１６　ディスクキャッシュ内容比較部
１７　比較結果通知部

【図１】



【図２】